Bias-Variance Tradeoff and Model Validation

Alexander McLain

August 20, 2025

Model Selection

Model Selection

The Problem of Model Selection

- In linear regression, we often have many potential predictors.
- **Model selection** = deciding which subset of variables to include.
- Why it matters:
  - Too few predictors $\Rightarrow$ underfitting
  - Too many predictors $\Rightarrow$ overfitting
- Goal: balance complexity and prediction accuracy

What Is "Learning Theory"?

- Statistical Learning Theory studies how algorithms trained on a sample perform on unseen data.
- Key question: **Will the model generalize?**

$$\text{Training error} \overset{?}{\approx} \text{Population error}$$

- Provides tools (bounds, capacity measures) to relate sample size, model complexity, and expected prediction error.
- We emphasize *intuition* over proofs; details become crucial only when you design new algorithms.

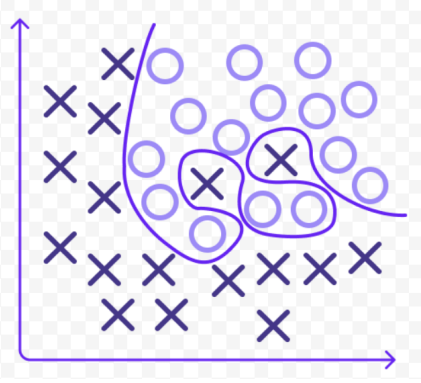
Memorization vs. Generalization

Memorization

- Perfectly fits training data (zero error).
- Fails catastrophically on new data.
- Example: lookup table of all training inputs.

Generalization

- Captures *structure* rather than noise.
- Training error is low *and* stays low on test data.
- Goal of any predictive model.

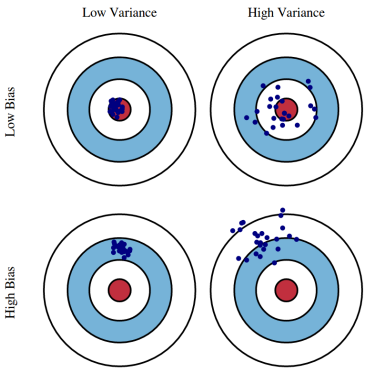


Example Scenario

- Predicting blood pressure using:
  - Age, weight, gender, cholesterol, physical activity, smoking, etc.
- Which of these predictors should we include in our model?
- What criteria can we use to choose?

The Bias–Variance Story

- **Bias**: systematic error from overly rigid models (under-fit).
- **Variance**: sensitivity to idiosyncrasies of training sample (over-fit).
- Sweet spot minimizes expected prediction error.
- Take-home: adding complexity helps until variance explodes.



Model Selection Goals:

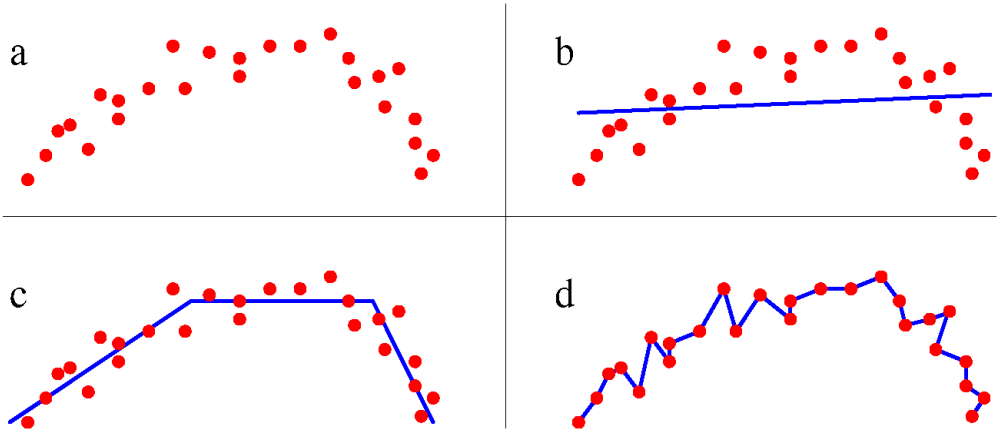
To make reasonable predictions or estimations, we need

- accuracy $\rightarrow$ on average, what we estimate is equal to what we expect (e.g., the predicted blood pressure is equal to the average blood pressure for all groups of people).
  - The predicted blood pressure equals the average blood pressure for all groups of people.
- precision $\rightarrow$ small variation in prediction/estimation
  - The predicted blood pressure is close to the true blood pressure for all groups of people.

How Do We Spot Overfitting?

- **Gap** between training and validation error grows.
- Highly erratic predictions for small data perturbations.
- Model relies on "spurious" variables (e.g., patient ID).
- Visualization: training curve vs. validation curve.

Underfitting vs. Overfitting



Underfitting vs. Overfitting

- Underfitting occurs when important regressors are left out of the model. Costs:
  - deficient models (i.e., missing patterns).
  - misinterpretations of variable relationships.
- Overfitting occurs when all important regressors are in the model, but some unimportant ones are, too. Costs:
  - unneeded complexity and increased variance of the predicted values.
  - widened confidence and prediction intervals.

Practical Tools for Better Generalization

Cross-Validation

- k-fold, LOOCV, stratified CV
- Provides unbiased estimate of out-of-sample error.

Regularization

- Penalize model complexity (ridge, lasso, dropout).
- Shrinks coefficients or removes them entirely.

Early Stopping

- Monitor validation loss during training (NNs, boosting).
- Stop when loss stops improving.

Ensembling

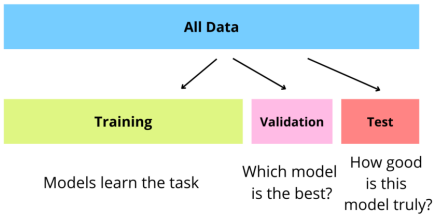
- Combine many weak learners (bagging, random forests).
- Averages out variance while retaining flexibility.

Evaluation Criteria

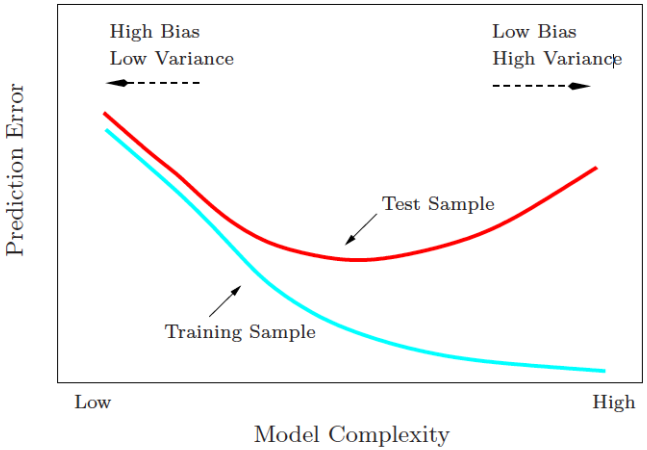
- **Training Error:** How well the model fits to the data used to estimate the parameters.
- **Test Error:** How well it predicts *test data*, i.e., data not used in the estimation of the parameters.
- **Cross-Validation:** procedure of cycling through the data, estimating the performance on disjoint groups of test data.
- **Information Criteria:**
  - AIC: balances fit and complexity
  - BIC: stronger penalty for complexity

Introduction to Validation

- Validation involves splitting the data into training, validation, and test sets.
- **Training Set:** The portion of data used to train the model, i.e., estimate $\hat{\beta}$.
- **Validation Set:** The portion of data used to tune model parameters and prevent overfitting, i.e., to tell us which subset of predictors to use in our model.
- **Test Set:** The portion of data used to assess the final performance of the model after training and validation.



Test vs Training Error



Introduction to Cross-Validation

- **Definition**: Cross-validation (CV) is a statistical method used to estimate the performance of machine learning models.
- **Purpose**: It helps in assessing how a model generalizes to an independent dataset.
- Why Use Cross-Validation?
  - **Prevents Overfitting:** Ensures model's robustness.
  - **Provides Reliable Estimates:** Offers a very good estimate of model performance.
  - **Optimizes Hyperparameters**: Helps choose parameters that can't be estimated (tuning parameters, e.g., the number of variables in a model).
  - **Accuracy Measures:** Any type can be used.

Basic Concepts

- **Divide Data:** Split data into training and validation (or test) sets.
- **Multiple Iterations:** Perform the split multiple times.
- **Aggregate Results:** Average the performance metrics.
- Types of CV:
  - K-Fold Cross-Validation
  - Leave-One-Out Cross-Validation (LOO)
  - Stratified K-Fold Cross-Validation
  - Time Series Split (for time-dependent data)

Cross-Validation Details

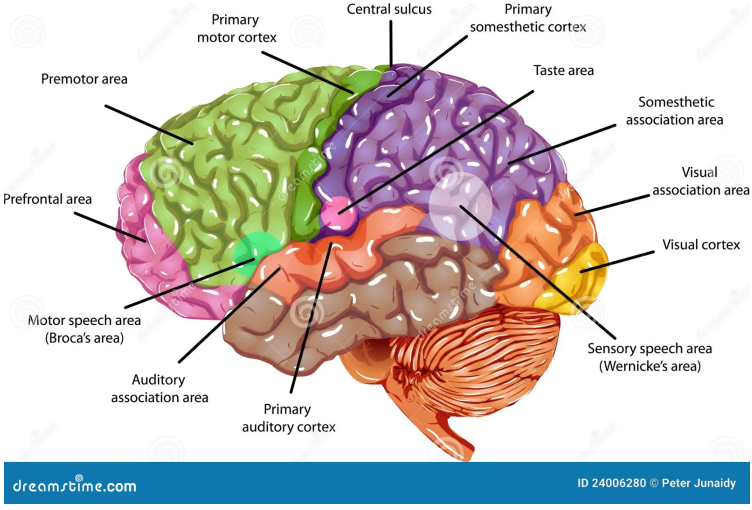
- Steps of *K*-fold cross-validation:
  1. Split data into K subsets (folds).
  2. Train/Estimate on K-1 folds and validate on the remaining fold.
  3. Repeat K times.
  4. Average the results.
- With LOO, each person is their own fold.

Multiple Linear Regression

- One dependent variable Y.
- p independent variables:
  - X_j = proportion of voxels damaged in Region Of Interest (ROI) j.
- The model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

  or $E(\boldsymbol{Y}) = \boldsymbol{X}'\boldsymbol{\beta}$.
- β_0: intercept;
- $\beta_j, j = 1, \cdots, p$: regression coefficients.
- $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2$


Primary
motor cortex
Central sulcus
Primary
somesthetic cortex
Premotor area
Taste area
Somesthetic
association area
Visual
association area
Prefrontal area
Visual cortex
Motor speech area
(Broca's area)
Sensory speech area
(Wernicke's area)
Auditory
association area
Primary
auditory cortex
dreamstime.com
ID 24006280 © Peter Junaidy

What $\boldsymbol{X}$ to use?

- Recall, that X_j = proportion of voxels damaged in ROI j.
- How do we define the regions? Some options:
  1. **Harvard-Oxford Atlas,** Number of ROIs (i.e., p): 48
  2. **Automated Anatomical Labeling (AAL) Atlas,** Number of ROIs: 116
  3. **Brainnetome Atlas,** Number of ROIs: 246
  4. **Schaefer Atlas,** Number of ROIs: Varies (100, 200, 300, 400, 500, 600, 700, 800, 900, 1000)
  5. **Voxel level,** $p > 10^5$.

Example: Predicting AQ with brain images (with 5-fold CV)

1. Put the people into 5 groups.
   Let $G_i = k$ if person i is in group k
2. Then for $k = 1, 2, 3, 4, 5$:
   - 2.1 Use $k - 1$ folds to fit a linear model to every brain atlas.
   - 2.2 Using the fitted linear models, predict AQ for every person in the kth group, i.e., the test fold.

   - 2.3 Get the total error for the test group for each brain atlas. For brain atlas 'a' this would be:

$$\widehat{SSE}_k^a = \sum_{i;G_i=k} (Y_i - \hat{Y}_i^a)^2$$

where $\hat{Y}_i^a$ is the predicted AQ when brain atlas 'a' is used.

Example: Predicting AQ with brain images (cont.)

3. Get the total error over all folds:

$$\widehat{SSE}^a = \sum_{k=1}^{5} \widehat{SSE}^a_k, \quad \widehat{SSE}^b = \sum_{k=1}^{5} \widehat{SSE}^b_k, \quad \widehat{SSE}^c = \sum_{k=1}^{5} \widehat{SSE}^c_k, \quad \text{etc.}$$

4. Use the ROI Options that has the lowest $\widehat{SSE}_j$.

This will give us an estimate of which ROI has the best predictive ability. No need to use forward/backward selection or model fit criteria.

Why the CV error of the "winner" is biased

- Suppose $\widehat{SSE}_a$ was the smallest.
- Think of CV error as $\widehat{SSE}_j = SSE_j + \varepsilon_j$ with $\mathbb{E}[\varepsilon_j] = 0$.
- However, we searched over many models and picked the $\varepsilon_{j^\star}$ with the best SSE.
- Selection makes $\varepsilon_{j^\star}$ tend to be **negative** (lucky noise).

$$\mathbb{E}\big[\widehat{SSE}_{j^\star}\big] \;<\; SSE_{j^\star}.$$

- This is the **winner's curse** / selection-induced optimism:
  - The more candidates you try, the larger the expected optimism.
  - Reusing the *same* CV folds for both tuning and reporting amplifies it.

How to get an honest performance estimate

- **Nested CV (recommended):**
  1. **Outer** split into K_{outer} folds.
  2. On each outer train set, run **inner** CV to tune/choose the model.
  3. Refit the chosen model on the full outer train set; evaluate on the outer hold-out fold.
  4. Aggregate outer-fold errors $\Rightarrow$ **unbiased** estimate for the selection procedure.
- **External test set:** If data permit, hold out a test set that is never touched until the very end.
- **One-standard-error rule:** Choose the *simplest* model whose CV error is within 1 SE of the minimum to reduce over-tuning.

Common Pitfalls

Common Pitfalls and Misuses

- **Data Leakage:** Ensure validation data do not influence training data.
  - e.g., centering the outcome and covariates at once.
- **Improper Splitting:** Avoid splitting time series data randomly (only want to predict forward).
- **Overlapping Data:** Ensure folds are mutually exclusive.

CV Takeaways

- **Proper Data Splitting:** Maintain clear boundaries between training, validation, and test sets.
- **Avoid Data Leakage:** Ensure no information from the test set influences the training process.
- **Avoid Data Overlap:** When working with grouped data (e.g., multiple visits per patient), ensure that groups are kept intact during cross-validation to prevent data leakage.
- **Respect the Temporal Order:** When working with time series or temporally ordered data, it is crucial to maintain the chronological order to avoid data leakage.
- **Evaluate Performance:** Use appropriate metrics and validation techniques to ensure reliable model performance assessment.

Google Flu Trends (2008–2014)

- **Idea**: Predict U.S. influenza levels by mining billions of Google search queries.
- **Early success**: Cross–validation on historical CDC data looked stellar.
- **Overfitting outcome**: Model keyed on ephemeral search spikes (media coverage, unrelated "flu" terms) and over-predicted the 2013 season by $\sim$140%.
- **Lesson**: Behavioral data streams drift quickly—continual re-training and ground-truth checks are essential.

Genomic "Sepsis Signature" in the ICU (2016)

- **Claim**: A 20-gene micro-array achieved 92% accuracy distinguishing bacterial vs. viral sepsis.
- **Problem**: Discovery and "validation" cohorts shared the same batch effects and platform.
- **Aftermath**: Independent hospitals saw accuracy drop to the low 60% range.
- **Lesson**: High-dimensional $p \gg n$ studies require external, multi-site validation and batch correction.

IBM Watson for Oncology (2015–2018)

- **Goal**: Recommend cancer treatments worldwide using cases from Memorial Sloan Kettering.
- **Issue**: Recommendations ignored local formularies, insurance rules, and rare pathology sub-types not seen in training.
- **Result**: Several hospitals withdrew after inconsistent or unsafe suggestions.
- **Lesson**: Elite single-site data can be parochial—robust generalization needs diverse, representative training sets.

Epic's Proprietary Sepsis Model (2017–2021)

- **Advertised performance**: AUROC ≈ 0.80 on internal tests.
- **Independent audit**: Three Midwestern hospitals found sensitivity $< 40\%$; many true sepsis cases missed.
- **Contributing factors**: Unknown feature weights, static thresholds, population shift.
- **Lesson**: Vendor black-box models are not exempt from overfitting—demand transparent, local validation.

COVID-19 Chest X-ray Classifiers (Spring 2020)

- **Competition hype**: CNNs reported $>95\%$ sensitivity distinguishing COVID pneumonia from "normal" lungs.
- **Confounder**: Datasets mixed adult COVID images with pediatric controls—models learned age-related anatomy, not pathology.
- **Reality check**: Balanced, age-matched test sets drove accuracy toward chance.
- **Lesson**: Spurious correlations lurk in hastily compiled data; stratify splits by site, scanner, and demographics.

Key Themes Across the Anecdotes

- Spurious signals and concept drift undermine apparent accuracy.
- Narrow, single-context training data rarely generalize broadly.
- Transparent external validation and continuous monitoring are non-negotiable safeguards.